Ο σχεδιασμός μιας ολοκληρωμένης μεθοδολογίας για την καθαριότητα δεδομένων περιλαμβάνει διάφορα στάδια και τεχνικές που συμβάλλουν στη διασφάλιση της ποιότητας των δεδομένων.

Το πρώτο βήμα θα περιλαμβάνει την **κατηγοριοποίηση όλων των πεδίων σε 2 βασικές κατηγορίες**, και συγκεκριμένα «Αλφαριθμητικά» και «Αριθμητικά» πεδία, καθώς η κάθε κατηγορία απαιτεί και διαφορετική προσέγγιση.

Μετά την κατηγοριοποίηση των πεδίων, πραγματοποιείται **εντοπισμός και αξιολόγηση των προβλημάτων που πιθανόν υπάρχουν στα δεδομένα** μέσω στατιστικών ελέγχων για την κατηγοριοποίηση και αξιολόγηση της εγκυρότητας των πεδίων. Όπως η παρουσία:

- λανθασμένων,
- ελλιπών ή
- διπλών εγγραφών.

Ο έλεγχος αυτός θα περιλαμβάνει (α) πίνακες συχνοτήτων, (β) ανάλυση κατανομών, (γ) έλεγχο διπλοτύπων και (δ) ποσοστό εμφάνισης αυτών.

Εν συνεχεία, γίνεται **εφαρμογή κανονικών εκφράσεων (regular expressions)** για τον έλεγχο των πεδίων. Η μεθοδολογία αυτή εφαρμόζεται τόσο για τα Αλφαριθμητικά , όσο και για τα Αριθμητικά πεδία.

Ενδεικτικά αναφέρεται, ότι με την χρήση κανονικών εκφράσεων (regular expressions) , είναι δυνατή η επικύρωση του αριθμού των ψηφίων ή λεκτικών (length validation) για τη διασφάλιση της συνέπειας των δεδομένων (π.χ. ΑΦΜ να διαθέτει 10 ψηφία). Επιπλέον με την ίδια μεθοδολογία γίνεται έλεγχος διαφορετικού λεξιλογίου σε ένα πεδίο, π.χ. αναγνώριση Ελληνικών και Λατινικών χαρακτήρων ταυτόχρονα σε ένα όνομα.

Στη συνέχεια, γίνεται **χρήση τεχνικών ασαφούς αντιστοίχισης** (fuzzy matching), όπως η απόσταση Levenshtein και ο **υπολογισμός του δείκτη ομοιότητας** (Similarity Score), για την αναγνώριση και ομαδοποίηση πολλαπλών εγγραφών που αντιστοιχούν στα ίδια λεκτικά. Η εφαρμογή αυτών των τεχνικών στις λεκτικές αναφορές των επωνυμιών επιτρέπει την ταυτοποίηση διαφορετικών παραλλαγών .

[περιγραφή για διασύνδεση με άλλες πηγές]

[πιο λεπτομερής περιγραφή ανά πεδίο / θα γίνει αφού ολοκληρώσουμε την καταγραφή του excel - συνάντηση με wemetrix]